Explorations into the social contexts of neologism use in early English correspondence

Säily, Tanja

2018

Säily , T , Mäkelä , E & Hämäläinen , M 2018 , ' Explorations into the social contexts of neologism use in early English correspondence ' , Pragmatics & Cognition , vol. 25 , no. 1 , pp. 30-49 . https://doi.org/10.1075/pc.18001.sai

http://hdl.handle.net/10138/304042
https://doi.org/10.1075/pc.18001.sai

acceptedVersion

Explorations into the social contexts of neologism use in early English correspondence

This is the 'author accepted manuscript' of the following paper: Säily, Tanja, Eetu Mäkelä & Mika Hämäläinen. 2018. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition* 25(1). 30–49. https://doi.org/10.1075/pc.18001.sai

Tanja Säily, Eetu Mäkelä and Mika Hämäläinen
University of Helsinki

1. Introduction

This paper describes ongoing work towards a rich analysis of the social contexts of neologism use in historical corpora, in particular the *Corpora of Early English Correspondence* (CEEC). The motivation for this work is to complement previous research on neologisms, which has tended to focus on lexicographical materials (see further e.g. Nevalainen 1999) or large present-day data (Renouf 2007; Grieve, Nini & Guo 2017). We present three case studies exploring neologisms in the eighteenth-century section of the CEEC: two focusing on the nominal suffixes *-ity* and *-er*, respectively, and one on a lexically creative individual, Thomas Twining.

Here, detecting historical neologisms is just a step in a larger process of sociolinguistic research, where we are more interested in the nuanced social and semantic aspects of their use. The types of research questions we are interested in are as follows:

1. Who are the innovators? Which social groups do they represent?
2. How do the new words spread socially, geographically and diachronically?
3. Which semantic domains do the neologisms represent?
4. Why are the neologisms created and established? What kinds of social meanings are associated with them?

To enable this kind of in-depth study of neologism use, new processes, tools and ways of combining data from different sources are required.

A significant data source, the *Oxford English Dictionary* (OED) provides us with information on the first attestations, etymologies and usage of words. However, the dictionary is admittedly biased, for instance, towards published texts by well-known authors (Brewer 2007; Hoffmann 2004). By contrast, the CEEC represents private writing produced by authors from a wide social spectrum, richly documented in the metadata accompanying the corpus – both essential factors in our pursuit of a more fine-grained understanding of the social embedding of neologism use. For semantic domain analysis, the *Historical Thesaurus* (HT; Kay et al. 2009) can be used to map words to semantic categories. Finally, we have at our disposal several databases of published texts with dates attached (e.g. *Early English Books Online* (EEBO), *Eighteenth Century Collections Online* (ECCO), *British Library Newspapers* (BLN)), against which neologism candidates can also be checked.

Combining all of these sources is, however, by no means straightforward. As far as our corpus is concerned, we are dealing with a collection of letters varying from the fifteenth century all the way to the early nineteenth century, including not only official and business correspondence but also informal correspondence between friends and family members. This means that we are faced with a tremendous amount of variation in orthography resulting from the lack of a common standard, different conventions in different eras, varying education levels of the writers, and simply from spelling mistakes.

Thus, in order to match the words in our corpus with the OED or HT entries, we have to perform two steps: one is to normalize the spelling variants to a standardized English spelling (e.g. from *refarred* to *referred*) and the other is to lemmatize the normalized word forms (e.g. from *referred* to *refer*). Finally, the text databases have mostly been automatically transcribed from microfilms, so in addition to spelling variation and inflection, they contain a large number of OCR errors which further complicate comparisons.

Our project brings together computer scientists and historical sociolinguists in trying to develop a process and an environment aimed at facilitating the historical-sociolinguistic study of neologisms, combining both qualitative and quantitative analysis.

Our method has been to start from concrete case studies that are focused enough to make the problems tractable. Through development of tools aiding these studies, we then incrementally build toward more general analyses as the supporting tools and data integrations mature. All in all, we are not aiming for a fully automatic pipeline to process such noisy data. Instead, we will develop an open-source environment where information on neologism candidates is gathered from a variety of algorithms and sources, pooled, and presented to a human evaluator for verification and exploration.

The case studies presented in this paper shed light on different aspects of the new lexis in the corpus and on the tools required. As nouns are the largest lexical category, we focus on nominal suffixes in our first two studies (Section 3). For diversity, we analyse an abstract noun suffix of a foreign origin (*-ity*) and a concrete noun suffix of a mixed origin (*-er*/*-or*). In our third study (Section 4), we widen our focus from a manual investigation of individual affixes to all words automatically matched to the OED so far, and highlight a person who emerged as the most innovative individual in the first two studies: Thomas Twining, a clergyman whose brother headed the Twining tea company and who was a peripheral member of eighteenth-century literary circles. By comparing the partial answers these studies provide to our research questions, we are able to not only gain important insights into the social variation in neologism use in the corpus (Sections 3 and 4), but also to assess the extent to which the corpus is capable of answering our questions, and to clarify the methodological issues for future work (Section 5).

2. *Corpora of Early English Correspondence* (CEEC)

Before proceeding to the case studies, a brief discussion of the corpus is in order. Based on published editions of letters, the CEEC (1402–1800) was designed for the purposes of historical sociolinguistics, and as such it aims at social representativeness in terms of gender and social rank. The speech-like genre of personal letters is uniquely well suited to the task, as letter-writers come from a much wider range of social backgrounds than writers of published texts. Nevertheless, the corpus is skewed

towards well-educated men of the upper ranks, as they were the most literate social group and it was mostly their letters which were considered important enough to be preserved and later edited.

In the long eighteenth century (1680–1800), on average a quarter of the running words in the CEEC was produced by female informants; by the final forty-year period, their share had risen to 33%. The proportion of letters by social rank is shown in Figure 1. If we compare the proportion of the middling ranks of merchants and professionals with that of the upper ranks of royalty, nobility and gentry, we can see that the share of the former increases so that they become roughly equal to the latter by the end of the century. The lowest rank of other non-gentry, too, reaches its highest values towards the end of the century. These changes reflect the increasing power and wealth of the middle classes, as well as increasing literacy rates (see further Kaislaniemi 2018).

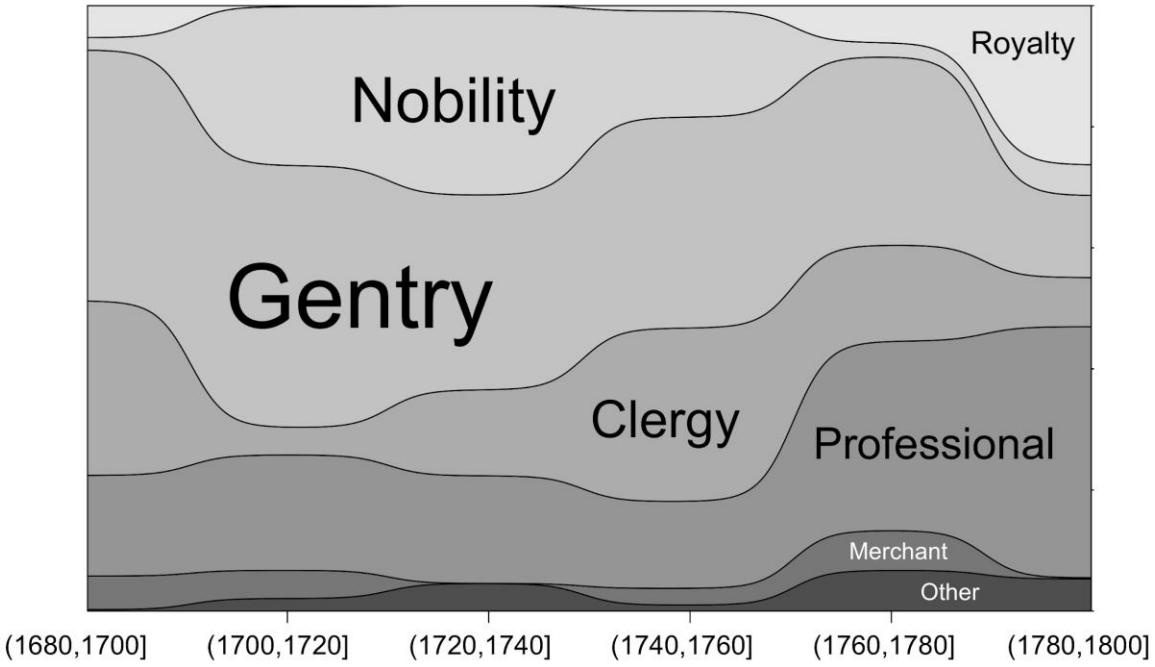


Figure 1. Density plot of the proportions of social ranks in the eighteenth-century section of the CEEC (courtesy of Harri Siirtola). Reproduced from Kaislaniemi (2018): Figure 4.4.

A register-related aspect of the composition of the corpus is the relationship between the sender and recipient of the letter. By the final forty-year period of the corpus, which we will focus on below, the proportions of running words by relationship are as follows: nuclear family 40%, other family 7%, close friends 32%, other acquaintances 21% (the total number of running words in the period is 1,051,564). The proportions within the social rank of professionals are similar: nuclear family 36%, other family 11%, close friends 36%, other acquaintances 17% (out of a total of 416,849 running words). Thus, the numbers are skewed towards nuclear family and close friends, but there is a non-negligible amount of data for the other relationships as well. For women, who were more confined to the private sphere, the proportion of family letters is somewhat higher than for men.

These facts about the corpus and eighteenth-century society at large should be kept in mind when evaluating the results of the case studies.

3. Two case studies of specific types of neologisms: *-ity* and *-er*

For our first case studies, we started with the in-depth investigation of a specific kind of neologisms – those formed by appending a particular suffix to a word. At this point, we did not have programmatic access to any of the other sources besides the CEEC, so our approach consisted of extracting all likely candidates for words formed in this way, and then developing an environment where the user could go through them as easily as possible to filter them and compare them with information in the web version of the OED.

For the two case studies presented in this section, our working definition of ‘neologism’ was that the corpus attestation should be no more than 100 years after the first attestation in the OED. This wider focus on recent vocabulary enabled us to gain a more comprehensive picture of the use of the suffix in question. In what follows, we will report all of the words found, but in the discussion of individual innovators, we will chiefly concern ourselves with the newest words.

3.1 Case study 1: *-ity*

Our first case study deals with the nominal suffix *-ity*, which is the Romance/Latinate counterpart of the native suffix *-ness*. These suffixes typically derive abstract nouns from adjectives, as in *generous* : *generousness* / *generosity*. As *-ity* was borrowed from French and Latin, it bears the connotations associated with those languages, such as 'polite society', 'learned' and 'scientific' (cf. Adamson 1989). Previous research (Säily 2014) has found that the productivity of *-ity* increased throughout the seventeenth and eighteenth centuries in the correspondence genre, perhaps led by male professionals. By the end of the eighteenth century, *-ity* was thus highly productive, so it is of interest to study what kinds of neologisms were produced during this period (1760–1800) and by whom.

Due to the large amount of spelling variation inherent in the corpus, finding all words created with the *-ity* suffix in the CEEC is not as simple as just searching for that suffix. Instead, through domain knowledge and experimenting, it was determined that the core unit of the spelling of *-ity* was one to two `t` characters towards the end of the word. This unit could be preceded by the characters `a`, `e`, `i`, `j`, `l`, `r`, `u`, `y`, or the abbreviation marker in the corpus, `~`. It could be followed by the characters `ee`, `ey`, `e ie`, `i`, `ye`, `y`, and optional plural/possessive forms. Thus, while the most common spelling was *-ity*, spellings like *-yttees*, *-uties* and *-jtty* were also possible. The possible spelling variants were charted for the entire time period covered by the corpus rather than just the eighteenth century, and the stranger variants were mostly associated with the earlier periods. However, even in the eighteenth century, less well educated writers could spell words quite variably, so in order to avoid missing instances from them, we decided to include the complete list of variants in our search.

For the long eighteenth century (1680–1800), this search yielded 24,053 occurrences (representing more than 1,900 different candidate types) in the letters. However, most of these were false hits – for example, words ending in *-ite* were quite common and typically not instances of *-ity*, although *-ity* could also sometimes be

spelled in this way. Using an Excel sheet and separate access to the CEEC, Säily (2014) pruned the hits down to 5,740 actual *-ity* tokens representing 297 types.

For the final forty-year period of 1760–1800, there were 264 *-ity* types. By manually retrieving the first attestation dates of each of these types through searches on the OED Online website, we found 15 neologisms in line with our definition in the CEEC, listed below. The three words in boldface antedate the OED (the years of first attestation in the corpus and the dictionary are given in parentheses), while the three that are underlined are not found in the dictionary at all.

> authenticity, **cleverality** (1778<1828), comicality, **conviviality** (1783<1791), coxcombicality, foxity (1788), **Germanity** (1788<1821), impracticability, intrepidity, irritability, oddity, respectability, ridiculability (1776), scoundrellity (1761), versatility

Looking at who produced these words, four men emerge as particularly innovative. Thomas Twining (c.1734–1804), clergyman and classical scholar, produced four of the words listed above – *ridiculability* (1776), *coxcombicality*, *foxity* and ***Germanity*** (all 1788) – although the meaning of *foxity* is unclear, and it is possible that he meant to write *Foxite* 'a follower of Charles James Fox' instead. In any case, he leads the group at three or four innovative words. Charles Burney (1726–1814), musician and author, was the father of author Frances Burney and a friend of Twining's. He produced *versatility* (1782) and ***conviviality*** (1783), the latter in a letter to his daughter, who used it back at him in 1799. David Garrick (1717–1779), actor and playwright, produced *scoundrellity* (1761) and *comicality* (1767). Ignatius Sancho (1729?–1780), author, produced ***cleverality*** (1778).

Many of the words seem to pertain to human attributes. Examples (1) and (2) illustrate how they were used by their coiners. We can see that these two words were used to refer to an undesirable quality or act attributed to members of an outgroup, which simultaneously serves to emphasize the ingroup membership of the writer and the recipient of the letter. While the words could have been coined in order to fill a perceived semantic gap, the choice of the etymologically foreign suffix *-ity* also carries

social meaning. It could signal the writer's learnedness (again strengthening ties with the recipient by assuming that they too would understand this learned word), but there also seems to be a humorous effect in this creative use of the suffix. All in all, the *-ity* words were used as part of an affective and involved, yet humorously elevated style in letters to close friends or family (see also Säily 2014:chap. 11).

(1) […] there is not among any set of people such a comfortable scratch-back confederacy as among those old ruin-diggers [antiquarians & old-English grubbers]. Is it the consciousness of enemies & scoffers all round them – the sour'd feeling of ridiculability – that draws the knot closer and forms the phalanx, back to back, for mutual scrubbing & defence? (Thomas Twining to Charles Burney, 1776; TWINING_019)

(2) It is a most infamous design, & I desir'd Churchill would Let Thornton know of it, which he will do immediatly, & prevent their Scoundrillity by some humourous Paragraph […] (David Garrick to George Colman, 1761; GARRICK_048)

Although most of these words are hapax legomena in our corpus, we can glean some evidence as to how neologisms might have spread from letters between Charles and Frances Burney (examples 3 and 4 below). While years have passed between the two instances of ***conviviality***, they illustrate how the use of a word may spread in a network of family and friends. Note that we do not have complete correspondences in our corpus, so that e.g. the letter to which Frances's letter is a response is not included, but it seems that she is repeating something her father wrote to her in a previous letter. She has also emphasized (probably through underlining in the original manuscript) the word *conviviality*, perhaps as a signal that she finds it slightly odd or novel (cf. Kaunisto 2013).

(3) I have always thought, that in many particulars his equal was not to be found – his wit, learning, taste, penetration; &, when well, his **conviviality**, pleasantry, – & kindness of heart to me & mine, will ever be thought of, with the most profound & desponding regret! (Charles Burney to Frances Burney, 1783?; BURNEY_031)

(4) The account of the Play meriting little attention indeed – I am much pleased at your independent establishment of ***conviviality*** at Burlington House. (Frances Burney to Charles Burney, 1799; BURNEYF_054)

This case study, then, suggests the following preliminary conclusions with regard to our research questions.

1. The innovators are creative men in their forties or fifties who belong to the social rank of professionals or other non-gentry.
2. The neologisms spread in a social network of peers writing in a similar style.
3. The words often describe human attributes.
4. In addition to their labelling function, the words carry social meaning: they are designed to amuse and perhaps impress the recipient and to emphasize ingroup membership.

3.2 Case study 2: *-er*

Our second case study concerns the nominal suffixes *-er* and *-or*, which typically derive agentive or instrumental nouns from verbs (e.g. *drive* : *driver*, *govern* : *governor*, *fill* : *filler*). They can also be added to nouns to denote a person 'concerned with N' or 'living in N' (e.g. *football* : *footballer*, *London* : *Londoner*). The suffix *-or* is the Latinate variant of *-er*; since they are pronounced identically (compare *adviser*/*advisor*), and since they were often spelled interchangeably, they are here treated as a single suffix (see Plag 2003:89; Bauer 2001:199–203). Ongoing research (Säily, Suomela & Mäkelä in preparation) has found that the productivity of *-er* increased over time in the long eighteenth-century section of the CEEC (1680–1800). Similarly to *-ity*, then, the suffix was highly productive by the end of the eighteenth century (1760–1800), so we would like to find out what kinds of neologisms were produced using it, and who produced them.

Again, due to the spelling variation in the CEEC, a quite complex search query needed to be constructed to find all candidate words. Here, the core unit of the spelling

of *-er* was determined to be an `r` character towards the end of the word. It could be preceded by anything and followed by an optional `e` character as well as optional plural/possessive forms. Thus, possible variants included for example *-er(e)*, *-ar(e)*, *-or(e)*, *-our(e)*, *-owr(e)* and *-ur(e)*, and their plural and possessive forms. Furthermore, there were a number of occupational *-er* words that could be abbreviated in a way that left no trace of the `r` character (e.g. *tailo~* for *tailor* and *carpunt~s* for *carpenters*); to capture these, we searched for the abbreviation marker preceded by `e` or `o`, or followed by plural/possessive forms.

In the end, we ended up with a total of 6,800 candidate types with over 300,000 appearances. With such a large number of words to verify and compare against the OED, we came to the conclusion that a dedicated tool for such work would be needed. Out of this came the FiCa tool, which is a user interface aimed at enabling the end user to, as quickly as possible, filter and categorize a set of data based on contextual information.

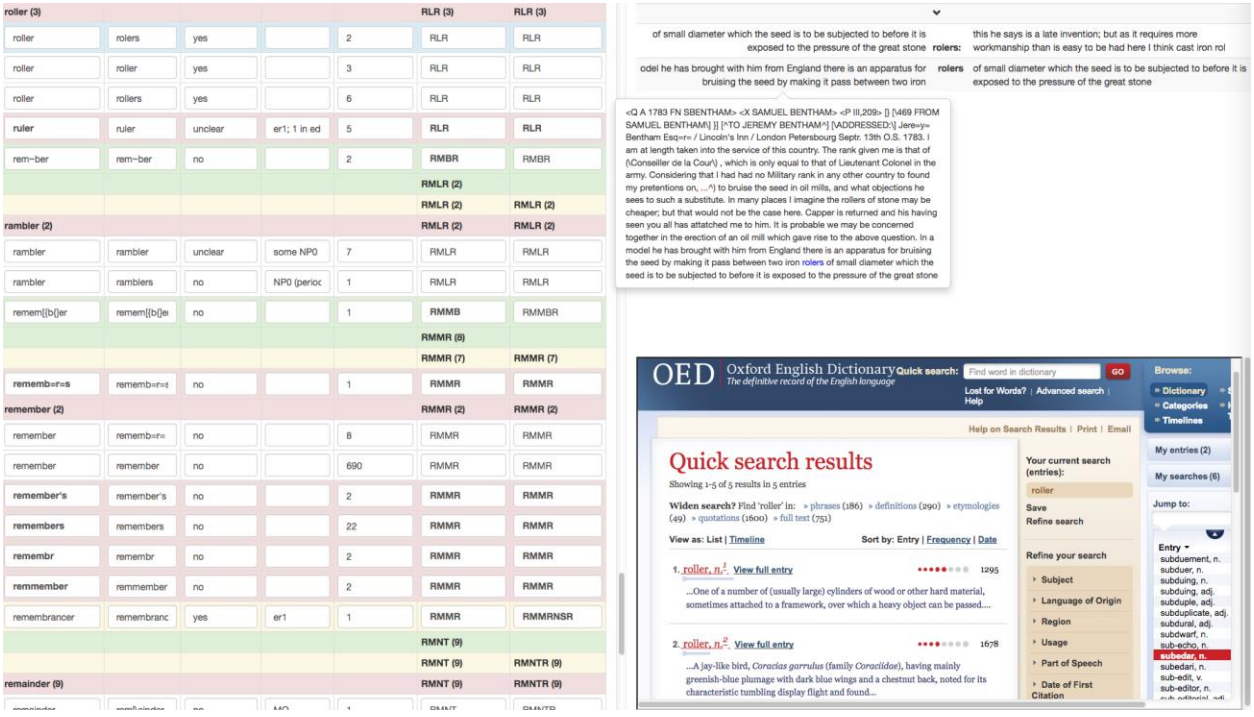
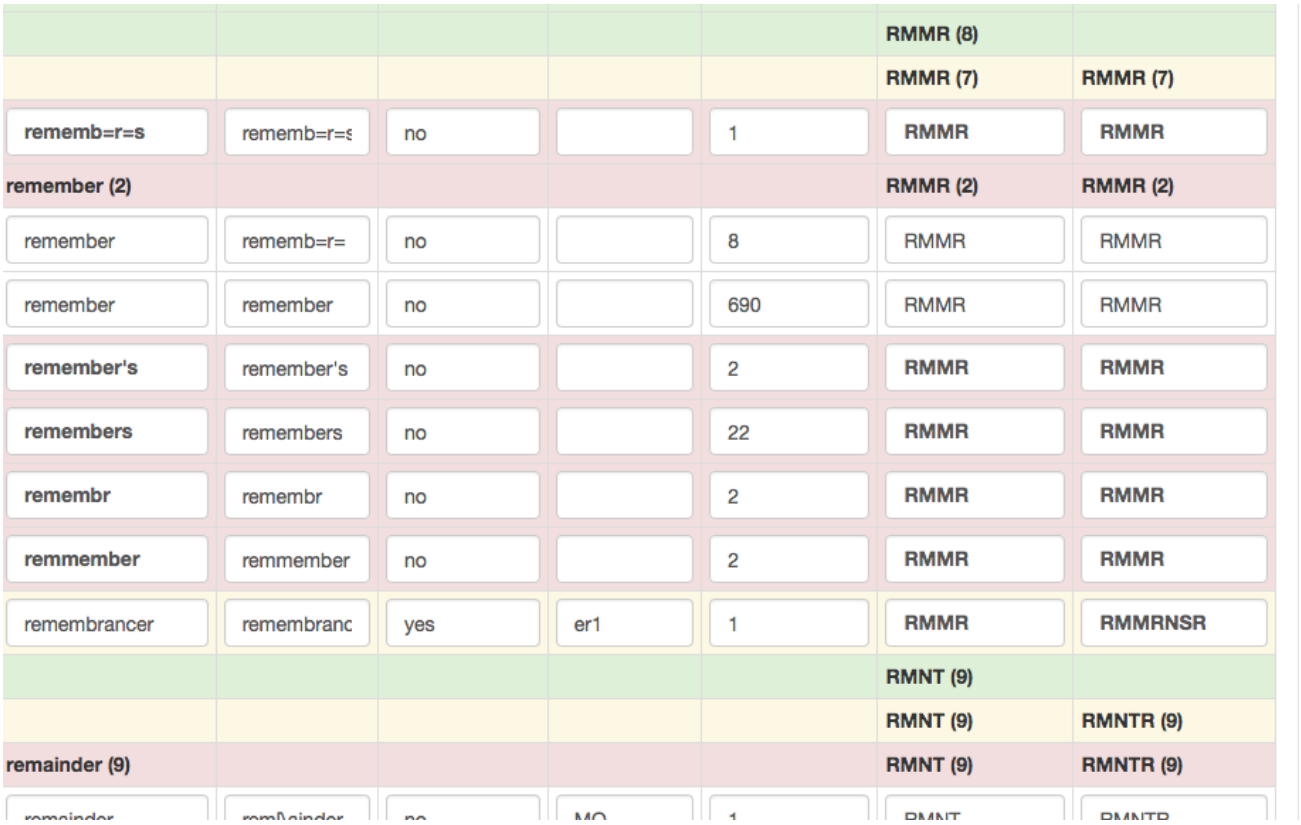


Figure 2. The FiCa interface.

The primary mode of working with FiCa is through the keyboard. In the FiCa interface, shown in Figure 2, the entries to process are shown in a spreadsheet-like interface on the left of the screen. Through hotkeys, the user is able to move through these, and make filtering and categorization decisions through quick keypresses. On the right, contextual information is automatically loaded for the row in focus to aid these decisions. In the configuration depicted, this information is, on the top right, how the word appears in its original letter context, and on the bottom right, the information on the word in the OED.

As said, in the CEEC, spelling variation is significant. This has relevance for filtering because out of the 6,800 surface forms, many are in fact the same word, just spelled differently. To make use of this fact, FiCa allows grouping of words hierarchically by algorithmically calculated keys, enabling the user to make decisions for a whole group using just a single keypress.

For the *-er* case, we used two main means of calculating such grouping keys. Firstly, we used VARD2 (Baron, Rayson & Archer 2009), which is a commonly used tool to assist in the normalization of the spelling of Early and Late Modern English texts. We have a VARD2 normalized, human validated version of a subset of our corpus, namely the letters from collections spanning the 16th–18th centuries. We used the VARD2 mapping between spelling variants and their normalizations in this version to extract normalized grouping keys for use in FiCa. However, in the VARD2 normalized version of CEEC, normalization was only applied to sufficiently frequent words, so a number of low-frequency types remain non-normalized (Palander-Collin & Hakala 2011). Thus secondly, as spelling variation in the absence of standardization often arises between phonetically similar forms, we made use of the Metaphone algorithm (Philips 2000) to calculate phonetic keys of two different granularities for the words.



					RMMR (8)	
					RMMR (7)	RMMR (7)
rememb=r=s	rememb=r=s	no		1	RMMR	RMMR
remember (2)					RMMR (2)	RMMR (2)
remember	rememb=r=	no		8	RMMR	RMMR
remember	remember	no		690	RMMR	RMMR
remember's	remember's	no		2	RMMR	RMMR
remembers	remembers	no		22	RMMR	RMMR
remembr	remembr	no		2	RMMR	RMMR
remmember	remmember	no		2	RMMR	RMMR
remembrancer	remembranc	yes	er1	1	RMMR	RMMRNSR
					RMNT (9)	
					RMNT (9)	RMNTR (9)
remainder (9)					RMNT (9)	RMNTR (9)

Figure 3. Hierarchical groupings in the FiCa interface.

In FiCa, these keys were then organized into a hierarchy, starting with the more general phonetic key, and progressing through the more fine-grained phonetic key and the normalized form to the individual words. Figure 3 shows an example of what such a grouping looks like in the spreadsheet portion of the interface. Here, the rightmost column is the max ten character Metaphone key, with a four character Metaphone key to its left. On the left side, on the other hand, are first the VARD2 normalized form of the word, and then the original form as it appears in the corpus. The middle columns are first the filtering and categorization columns, and finally a frequency column for information. In the interface, groups are shown as uneditable rows, highlighting the common row values that make up the group key.

Here, the highest level group line at the top shows that the eight lines below it are grouped under the Metaphone4 key of **RMMR**. Seven of those are further grouped also under the exact same Metaphone10, while the final one has a longer Metaphone10 key of **RMMRNSR** (note that distinct hierarchical group rows are not shown where they do

not add any information or distinct choices, as here). Inside the Metaphone10 groups, all entries are themselves distinct apart from one instance, where the VARD2 based normalizer has determined that *rememb=r=* (== here encoding a superscript) also normalizes to *remember*.

Using FiCa this way, none of the algorithms are trusted blindly, but similar words are grouped together regardless of their exact spelling. If the words in a higher-level group happen to represent the same lemma, decisions can be made on that level using a single keypress. In this case, the Metaphone10 groups are the level of meaningful distinction, so the right course of action is to filter out the group with the key of **RMMR** and keep the one line with *remembrancer* and the **RMMRNSR** key, as well as encode the *-er1* (agentive) category for this word, determined from the CEEC and OED context views. When the algorithms do not find suitable lemma forms for the word, FiCa also allows the user to change the generated form, thus verifying and correcting the output and grouping. This also automatically updates the OED context view to show the proper word for making grounded decisions.

In the end, a vast majority of 5,080 types out of the 6,800 were deemed irrelevant. On the other hand, 153 types were identified to contain homonyms, where certain individual uses needed to be counted while others needed to be discarded. These yielded a further 11,768 individual uses for inspection. The final number of *-er* types identified was 639, out of which 456 occurred in the period 1760–1800.

Out of this more general set of *-er* words, we then used the OED view in FiCa to identify neologisms. As in the previous case study, we looked for words that occurred in the corpus at most 100 years after their first attestation date in the OED. Using FiCa, we discovered 25 *-er* words meeting this criterion in the CEEC; they are listed in Table 1, loosely grouped by their meanings. The words in boldface either antedate the OED or represent the same instance in both the OED and the corpus (with the years of first attestation in the corpus and the dictionary given in parentheses), while the underlined words (all place-related) are not found in the OED at all. The place-related nouns could have been deliberately left out of the OED as rare and quite transparent formations, so we shall not put too much weight on them, although *Norfolker* ‘Norfolk cow’ seems to be more term-like than the rest.

Describing people	absconder, **blubberer** (1782<1786), **commemorator** (1784<1856), completer, complimenter, dangler, **outsider** (1800=OED), schemer, seceder, **spiter** (1790<1847), swindler
– *Occupations*	(shirt-)airer, gambler, hairdresser, (China-)piecer, smuggler
Connected to places	Chiswicker, Madrasser, Norfolker, Turnham-Greener
Things	cutter ‘boat’, ventilator
Other	brightener, **plumper** ‘lie’ (1776=OED), **winterer** ‘animal kept over the winter’ (1784<1795)

Table 1. *-er* neologisms in the CEEC, 1760–1800.

The list of most innovative letter-writers (who produced at least one antedating) is again headed by Thomas Twining (c.1734–1804), clergyman and classical scholar. He produced ***plumper*** (1776), ***blubberer*** (1782), ***commemorator*** (1784) and *complimenter* (1788). However, the rest of the innovators – or early adopters – are new: agriculturist George Culley (c.1735–1813), with the cattle-related words *Norfolker* and ***winterer*** (both 1784); Hester Piozzi (1741–1821), writer and Bluestocking, who produced *seceder* and ***spiter*** (1790); and novelist Jane Austen (1775–1817), with the first recorded use of ***outsider*** (1800). Notably, the innovators include both men and women, and the presence of Culley shows that it was not only professional writers who engaged in lexical innovation.

The neologisms in *-er* seem to display greater semantic variation than those in *-ity*. As noted by an anonymous reviewer, this could be due to the fact that there are more results for *-er* than for *-ity*; another possible contributing factor is that *-er* is overall a more heterogeneous word-formation process, as discussed at the beginning of this

section. Some of the meanings of the new formations stray quite far from the prototypical agentive/instrumental core of *-er*. Nevertheless, senses to do with people are again the most common. These are illustrated in examples (5) and (6), the latter of which includes a similar negative reference to an outgroup member as those in (1) and (2) above. Example (5), on the other hand, is humorously self-deprecating: "Cecilia" refers to Burney's daughter's novel, which made Twining cry as it had done to "two amiable sisters in Colchester, sensible & accomplished women, who were found blubbering at such a rate one morning", as he explains earlier in the letter. The point of the expression is thus to compliment Burney's daughter on writing such a moving novel. Both of the examples can be said to reflect an involved style of writing designed to strengthen ties with the recipient.

(5) As to myself, Cecilia has done just what she pleas'd with me: I laughed, & cried (for I am one of the **blubberers**) when she bade me. (Thomas Twining to Charles Burney, 1782; TWINING_033)

(6) it makes me laugh when I think how the **Spiters** told us that Siddons had lost all her Popularity […] (Hester Piozzi to Charlotte Lewis, 1790; PIOZZI_027)

Owing to the fact that all of the *-er* antedatings are hapax legomena in the corpus, we have no evidence for how they spread. However, as was the case with *-ity*, all of them occur in letters written to close friends or family, so it is plausible that they would have spread in these social networks of the innovators. Of course, not all of them may have spread, and the later attestations (e.g. in the OED) could have been recoined. Whether or not a neologism caught on could in part depend on its user's status in the network. For instance, while Thomas Twining was a highly creative language user, he was probably not one of the most central and influential members of his network, whose adoption of a word would have ensured its diffusion (cf. Conde-Silvestre 2012).

Our second case study offers the following preliminary insights into our research questions.

1. The innovators are both men and women, mostly in their forties or fifties, who belong to the social rank of professionals or other non-gentry.
2. If at all, the neologisms spread in a social network of peers writing in a similar style.
3. They often describe people.
4. They also carry social meaning: many of them seem to be nonce-formations designed to amuse the recipient and to emphasize ingroup membership. Some reflect innovations in society or (agri)culture and were more obviously coined to fill a semantic gap.

In sum, the two case studies would seem to indicate that the late eighteenth-century neologisms in the CEEC were mostly produced by the emerging middle class and that they reflected an involved, affective and interpersonal style of letter-writing. However, the most innovative social groups vary by affix: *-ity*, with its learned connotations, was chiefly used by men in creative professions, while *-er* was used by both men and women, both professionals and other non-gentry. Moreover, the purpose of the innovation may vary by affix, as *-er* was also used for societal innovations. The fact that we found (male) professionals writing to close friends and family members to be consistently among the most innovative groups is perhaps unsurprising, as they are the largest social group in the corpus. Nevertheless, the example of *-er* shows that it is possible to gain results regarding other social groups as well, and in a larger sample the differences in subcorpus size could be accounted for using statistical methods (e.g. Säily & Suomela 2017). It is therefore clear that we need to go beyond case studies and aim at a broader analysis.

4. Towards computational discovery of neologisms in general

Expanding the analysis from individual means of neologism creation to all neologisms in the CEEC manually would be an immensely time-consuming and laborious undertaking. Therefore, since obtaining computational access to the other materials, we have been

exploring a more automated method of extracting potential neologism candidates from the corpus. Extracting candidates narrows down the number of words we have to go through in the corpus to study neologisms, reducing the time it takes to conduct qualitative research on them.

With full access to the OED in an XML format (thanks to an agreement with Oxford University Press), we can now automatically compare each word in the CEEC with the earliest attestation marked in the OED. If a word has appeared in a letter that has been written earlier than the earliest attestation recorded in the OED, we assume we are dealing with a neologism candidate.

However, the problems of spelling variation and lemmatization still remain. In our current process, when we iterate over all the words in the corpus, we will first try to lemmatize each word. For lemmatization, we use the NLTK Python library (Bird, Klein & Loper 2009), which provides a lemmatizer based on WordNet (Miller 1995). This NLTK function is given a lowercase word form as input and it produces a lemmatized form as output. We can assess whether the lemmatization produced a lemma by checking if the output exists in the OED as a lemma. If the lemmatizer fails, we will try to normalize the word form by more specific methods.

First, we try the VARD2 normalized form substitutions mentioned earlier. The OED also has a list of alternative spellings and inflections for its entries. We use this data as well to directly lemmatize spelling variants. Finally, if all the previous options have failed, we try processing the word with MorphAdorner (Burns 2013), which is another tool developed for normalizing early English texts. It is important to note that all of these approaches match words against the OED only on a string level, i.e. the automated approach does not take higher-level linguistic variation such as homonymy or polysemy into account.

With these methods, we have successfully obtained the earliest attestation from the OED for 65,848 word forms appearing in our corpus, but we still have 85,362 word forms we have been unable to lemmatize or normalize. This calls for more research in that vein, and we are currently looking into developing our own methods for old text normalization to increase the coverage in our corpus.

4.1 Case study 3: Thomas Twining

Using the lemmas automatically matched to the OED so far, we have already been able to discover interesting neologisms. We decided to focus on the eighteenth-century clergyman Thomas Twining, who was identified as a creative language user in the previous case studies of *-ity* and *-er*. By filtering the results to letters written by Twining and by sorting them by the difference between the first attestation in the letter vs. the OED, we quickly came up with a list of 19 words produced by Twining that either antedate the OED or that are the OED's first attestation. These are presented loosely grouped by their meanings in Table 2 (the years of first attestation in the corpus and the dictionary are given in parentheses). In future research we will use the HT to classify the words more precisely.

People	**blubberer** (1782<1786), **commemorator** (1784<1856), **moderationist** (1792=OED)
– People's actions	**moon** (verb, 1763=OED), **moonery** (1764<1834), **pushery** (1788=OED), **truantism** (1785<1812)
– People's attributes	**embodiment** (1777<1828), **inside-outness** (1788<1919), **scratch-back** (1776<1842)
Music	**bravura** (1783<1787), **crescendo** (1773<1776), **keyless** (1781<1816), **monotonous** (1774<1776)
Writing	**conspectus** (1788<1839), **jargonic** (1781<1794), **letteret** (1799=OED)
Other	**jumpable** (1765<1829), **slushing** (1775<1863)

Table 2. Neologisms produced by Thomas Twining in the CEEC, 1763–1799.

Note that we are missing some words that were identified in the previous, more manual case studies. ***Germanity*** was not identified because the words are lowercased in the lemmatization process, and the word *germanity* in the OED is earlier than the *Germanity* used by Twining. We are also not yet considering semantic neologisms, so ***plumper*** in the sense 'lie' was not discovered. ***Moon*** as a verb was only found because it occurred in the form *mooning*, which was falsely matched to the noun *mooning*. Moreover, words not in the OED are of course not found in this process, so e.g. *ridiculability* is missing.

Nevertheless, the collection of neologisms in Table 2 is indicative of Twining's lively and involved style, with a great number of references to people. The collection is also representative of the semantic domains to which he contributed, which include his personal and professional interests, music and writing. In addition, we are able to get a sense of the impressive range of word-formation processes he used: borrowing from Italian and Latin (***bravura***, ***crescendo***, ***conspectus***), derivation with both native and etymologically foreign suffixes (*-er*, *-ing*, *-less*, *-ness*, *-able*, *-ery*, *-et*, *-ic*, *-ism*, *-ist*, *-ment*, *-ous*) as well as zero derivation (***moon***) and compounding (***scratch-back***). While any native speaker might be expected to have the ability to use any productive word-formation process in the language, it is clear that not everyone makes use of them to this extent and this creatively (cf. the previous two case studies, in which Twining was identified as the top innovator).

Two pieces of evidence illustrate how these words may have spread. Firstly, ***bravura*** (examples 7 and 8) was used by Twining's friend, Charles Burney, two years earlier than Twining (1783 vs. 1785), but as an instance of code-switching in the Italian phrase *mezza bravura*. As Italian was the language of music, the term must have been in the air otherwise as well, and it was only a matter of time before it would be used within English; the social network of the musically inclined Twining and Burney seems to have been among the first to do so.

(7) His voice was in good order, & he sang divinely – his 1st song a Grasiosa – or rather *mezza* ***bravura*** Air of Bertoni, which I had never heard before; but so elegant & fanciful that I shd have been less surprised had I been told that the

Catilena was Pacchierotti's – his 2^{d} Air a Cantabile in the *gran gusto* by Anfossi – and the 3^{d} air, *Rasserena il mesto ciglio*, of Gluck, w^{ch} we got encored. (Charles Burney to his daughter Susanna Phillips, 1783?; BURNEY_036)

(8) She sung, to our sorrow, only two songs; one, "I know that my Redeemer", &c., & a **bravura** song of – I forget – but it was a wonderful exhibition of power of voice, compass, distinct rapidity, & everything that cou'd be in such a song. (Thomas Twining to Charles Burney, 1785; TWINING_040)

The second piece of evidence concerns the word ***monotonous***. The third edition of the OED attributes *monotonous* to the Burneys: the literal sense 'having little or no variation in tone, pitch, or cadence' to Charles Burney (1776), and the extended sense '(anything) lacking in variety' to Frances Burney (1780). Twining used *monotonous* in the literal sense in a letter to Charles Burney in 1774, so the word seems to have spread in their social network (see examples 9–11 below).

(9) How I shall accent & express, after having been so long cramped with the **monotonous** impotence of a harpsichord! (Thomas Twining to Charles Burney, 1774; TWINING_017)

(10) A sound so much the more agreeable, as it is not **monotonous**, which is the case in the warble of most other birds. (Charles Burney, *A General History of Music*, 1776; OED)

(11) His lady – tittle-tattling, **monotonous**, and tiresome. (Frances Burney, *Diary and Letters of Madame d'Arblay*, 1780; OED)

5. Discussion

In this paper, we have described our recent work on delving in-depth into the contexts of neologism use in early English. As this is a report on work in progress, there are still multiple avenues for improvement.

First, using the OED alone as a yardstick for neologisms is naturally fraught with problems. As an example, many of our antedatings pertain to entries from the second edition of the OED that have not been updated since the late nineteenth century. These antedatings are therefore perhaps unlikely to represent genuine first attestations but simply gaps in the record of the OED. In general, it is doubtful whether we are ever able to gain access to the first use of a word from which it starts to spread – for instance, present-day scholars are missing most of the spoken language of the previous centuries. Nevertheless, we would argue that by using the materials that do exist we can get fairly close, even if who we call 'innovators' above should in many cases be called 'early adopters' instead.

Now that we have acquired full local access to large text databases of published English, we hope to be able to use them as additional evidence against which to weigh our neologism candidates, as well as material in which to track their diffusion. Here, comparing against ECCO and BLN, our preliminary results show that of all the CEEC types, some 36,000 have fewer than 100 appearances in the comparison corpora, c. 17,000 have fewer than 10, c. 2,500 appear once, while c. 6,700 do not appear at all. However, this preliminary analysis does not account for either spelling variation, OCR errors or inflection. Furthermore, as much of our correspondence predates the more standardized English in ECCO and BLN, we will at least need to add EEBO to the comparison corpora before doing any more thorough analysis of these.

On the other hand, the CEEC is the only one of our corpora where we have detailed social metadata on the authors, required for a nuanced sociolinguistic analysis. To counteract the relatively small size of the CEEC (5.3 million words spread across four centuries), which makes tracking individual neologisms difficult, we are experimenting with using the *Historical Thesaurus* to map words to broader semantic categories. Similarly to Marc Alexander and Christian Kay (2014), who used the HT to track the development of a semantic category, we could analyse recent vocabulary pertaining to a category in our corpus, but focus on the varying extent to which different social groups adopted the vocabulary over time (e.g. which social groups led the change).

From a computational viewpoint, in this paper we presented our initial approach to solve the problem of normalizing letters written in earlier varieties of English with spelling variation. As we pointed out earlier, there is still room for improvement in order to achieve a high coverage in the data. Previous research has suggested several automated methods for normalization that could be incorporated in our approach, such as comparing non-normalized words with the normalized ones by edit distance and distributional semantics (Amoia & Martinez 2013) or training a statistical character-level machine translation model to translate non-normalized word forms to normalized ones (Scherrer & Erjavec 2016). However, based on our preliminary trials with the CEEC, these methods, while also delivering correct results, produce a significant amount of noise, thus reducing precision.

As for recall, the case study in Section 4.1 highlighted some of the problems with relying on an approach based on pure string matching, such as disregarding homonymy and zero derivation. Moreover, attempting to match all of the words in a corpus to a dictionary is ultimately futile, as for instance most proper nouns will not be listed in a dictionary. Since part of the CEEC has been POS-tagged (PCEEC 2006), and the tagging has been checked manually, we could use the tagging to distinguish between e.g. nouns and verbs and to identify proper nouns. Again, however, this would only solve part of the problem, as the tagging does not cover the entire corpus and it was done without lemmatization. The tagging would also not help with issues such as identifying compounds written separately.

Thus, in the future, we will still need to abstain from the idea of finding a single silver bullet to solve the problem of normalization as a whole, but rather, continue to strive towards a hybrid approach in which the results of multiple different automated approaches could be compared and verified with as little effort as possible by the scholarly end user. While this kind of data processing and cleanup is likely to remain the most labour-intensive stage of the research process, our goal is to enable scholars to move past it more quickly to focus on exploration, analysis and interpretation.

Acknowledgements

We would like to thank the participants of the DynLex workshop for their helpful feedback. We are also grateful to the anonymous reviewers for their insightful input, which greatly improved this paper. Many thanks are due to the members of the STRATAS project, particularly Anna Merikallio, for useful comments and suggestions. This work was supported in part by the Academy of Finland, Grants 293009 and 276349.

References

Adamson, Sylvia. 1989. With double tongue: Diglossia, stylistics and the teaching of English. In Mick Short (ed.), *Reading, analysing and teaching literature*, 204–240. London: Longman.

Alexander, Marc & Christian Kay. 2014. The spread of RED in the Historical Thesaurus of English. In Wendy Anderson, Carole P. Biggam, Carole Hough & Christian Kay (eds.), *Colour studies: A broad spectrum*, 126–139. Amsterdam: John Benjamins.

Amoia, Marilisa & Jose Manuel Martinez. 2013. Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In Piroska Lendvai & Kalliopi Zervanou (eds.), *Proceedings of the 7th workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2013)*, 84–89. Stroudsburg, PA: Association for Computational Linguistics.

Baron, Alistair, Paul Rayson & Dawn Archer. 2009. Automatic standardization of spelling for historical text mining. In Claire Warwick (ed.), *Digital Humanities 2009: Conference abstracts*, 309–312. College Park, MD: Maryland Institute for Technology in the Humanities.

Bauer, Laurie. 2001. *Morphological productivity* (Cambridge Studies in Linguistics 95).

Cambridge: Cambridge University Press.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.

Brewer, Charlotte. 2007. *Treasure-house of the language: The living OED*. New Haven: Yale University Press.

Burns, Philip R. 2013. *Morphadorner v2: A java library for the morphological adornment of English language texts*. Evanston, IL: Northwestern University. http://morphadorner.northwestern.edu/morphadorner/. (19 May, 2018.)

CEEC. *Corpora of Early English Correspondence*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg et al. at the Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/. (19 May, 2018.)

Conde-Silvestre, Juan Camilo. 2012. The role of social networks and mobility in diachronic sociolinguistics. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The handbook of historical sociolinguistics* (Blackwell Handbooks in Linguistics), 332–352. Chichester: Wiley-Blackwell.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21(1). 99–127.

Hoffmann, Sebastian. 2004. Using the OED quotations database as a corpus – a linguistic appraisal. *ICAME Journal* 28. 17–30.

Kaislaniemi, Samuli. 2018 (in press). The *Corpus of Early English Correspondence Extension* (CEECE). In Terttu Nevalainen, Minna Palander-Collin & Tanja Säily (eds.), *Patterns of change in 18th-century English: A sociolinguistic approach* (Advances in Historical Sociolinguistics 8). Amsterdam: John Benjamins.

Kaunisto, Mark. 2013. Scare quotes and glosses: Indicators of lexical innovation with affixed derivatives. In R. W. McConchie, Teo Juvonen, Mark Kaunisto, Minna Nevala & Jukka Tyrkkö (eds.), *Selected proceedings of the 2012 symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*, 97–106. Somerville, MA: Cascadilla Proceedings Project.

Kay, Christian, Jane Roberts, Michael Samuels & Irené Wotherspoon (eds.). 2009.

Historical Thesaurus of the Oxford English Dictionary. OED Online. Oxford University Press. http://www.oed.com/thesaurus. (19 May, 2018.)

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

Nevalainen, Terttu. 1999. Early Modern English lexis and semantics. In Roger Lass (ed.), *The Cambridge history of the English language, III: 1476–1776*, 332–458. Cambridge: Cambridge University Press.

OED. *Oxford English Dictionary*. OED Online. Oxford University Press. http://www.oed.com. (19 May, 2018.)

Palander-Collin, Minna & Mikko Hakala. 2011. Standardized versions of the Corpora of Early English Correspondence. *Corpus Resource Database (CoRD)*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/standardized.html. (19 May, 2018.)

PCEEC. 2006. *Parsed Corpus of Early English Correspondence, tagged version*. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/. (19 May, 2018.)

Philips, Lawrence. 2000. The double metaphone search algorithm. *C/C++ Users Journal* 18(6). 38–43.

Plag, Ingo. 2003. *Word-formation in English* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Renouf, Antoinette. 2007. Tracing lexical productivity and creativity in the British media: 'The chavs and the chav-nots'. In Judith Munat (ed.), *Lexical creativity, texts and contexts* (Studies in Functional and Structural Linguistics 58), 61–89. Amsterdam: John Benjamins.

Säily, Tanja. 2014. *Sociolinguistic variation in English derivational productivity: Studies and methods in diachronic corpus linguistics* (Mémoires de la Société Néophilologique de Helsinki XCIV). Helsinki: Société Néophilologique.

Säily, Tanja & Jukka Suomela. 2017. *types*2: Exploring word-frequency differences in corpora. In Turo Hiltunen, Joe McVeigh & Tanja Säily (eds.), *Big and rich data in*

English corpus linguistics: Methods and explorations (Studies in Variation, Contacts and Change in English 19). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/19/saily_suomela/. (19 May, 2018.)

Säily, Tanja, Jukka Suomela & Eetu Mäkelä. In preparation. Variation in morphological productivity in the history of English: The case of *-er*.

Scherrer, Yves & Tomaž Erjavec. 2016. Modernising historical Slovene words. *Natural Language Engineering* 22(6). 881–905.

Authors' addresses

Tanja Säily
tanja.saily@helsinki.fi

Eetu Mäkelä
eetu.makela@helsinki.fi

Mika Hämäläinen
mika.hamalainen@helsinki.fi

P.O. Box 24
FI-00014 University of Helsinki
Finland

About the authors

Tanja Säily is a Postdoctoral Researcher in the Department of Digital Humanities at the University of Helsinki. Her research interests include corpus linguistics, digital humanities, historical sociolinguistics, and English lexis and morphology. In collaboration with computer scientists, she has developed several corpus-linguistic tools

and methods, and she is a co-compiler of the *Corpora of Early English Correspondence*.

Eetu Mäkelä is a tenure track professor of Humanities–Computing Interaction at the University of Helsinki, and a docent (adjunct professor) in computer science at Aalto University. His field is in developing computational tools to support humanities and social science research based on rich unstructured and structured data.

Mika Hämäläinen is a PhD student in language technology in the Department of Digital Humanities at the University of Helsinki. He has an MA degree in Spanish philology and his current research focuses on NLP for low-resource languages. He also has a background in computational creativity and NLG for morphologically rich languages.